University of Groningen

Off-line answer extraction for Question Answering

Mur, Jori

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2008

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):
Mur, J. (2008). *Off-line answer extraction for Question Answering*. [Thesis fully internal (DIV), Rijksuniversiteit Groningen]. s.n.

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 04-05-2023

Chapter 6

Conclusions

In this last chapter we will first summarise the main findings of this thesis. Then we will give suggestions for further research.

6.1 Summary of main findings

The theme of this thesis is question answering (QA), and its focus lies on finding answers to questions off-line. The addition of off-line answer extraction to an online question answering system is not only useful in terms of speeding up the process, the techniques applied can also differ, taking advantage of different things, that otherwise would remain unfeasible, such as extracting answers from the entire corpus, rather than just from the passages retrieved by the IR engine. Furthermore, we can filter out noise using frequency counts and classify answers off-line.

After a general introduction into question answering in chapter 1, we argued that knowledge about the possible realisations of dependency relations in questions and answer candidates can help to make the performance of QA more effective.

Off-line answer extraction is based upon the observation that certain answer types occur in fixed patterns, and the corpus can be searched off-line for these kind of answers. A problem associated with using patterns to extract semantic information from text is that patterns typically yield only a small subset of the information present in a text collection, i.e. the problem of low recall.

The aim of the present thesis was to address the recall problem by using extraction methods that are linguistically more sophisticated than simple surface pattern matching techniques. Specifically, the questions we have tried to answer in this thesis are the following:

- How can we increase the coverage of off-line answer extraction techniques without loss of precision?
- What can we achieve by using syntactic information for off-line answer extraction?

To answer these questions we performed a number of experiments discussed in the previous chapters. We have built an answer extraction module, Qatar, which is part of a Dutch state-of-the-art question-answering system called Joost. In order to have syntactic information at our disposal we parsed both corpus and questions with Alpino, a wide-coverage dependency parser for Dutch. Most questions used in our experiments were drawn from the CLEF question sets from 2003 to 2006. For some experiments in chapter 5 we created our own questions. The corpus in which we tried to find answers to these questions is the CLEF newspaper corpus for Dutch.

Even though all experiments use Dutch tools to find Dutch answers to Dutch questions, we have no reason to believe that the results of our experiments do not also apply to other languages.

In chapter 2 we performed a small experiment in which we automatically answered function questions. Analysing the results we showed the main problems with the technique of off-line question-answering. First, it turned out that the patterns used in the experiment were not robust enough. A slight difference in the sentence caused a mismatch with pre-defined patterns. Second, some questions contained important modifiers. These questions are hard to account for by off-line methods. Furthermore, we showed that some questions would have been answered if reference resolution techniques were implemented. Finally, sometimes the right pattern was simply not defined or it could have been better defined. We concluded that answers to questions in a natural language corpus are often formulated in such complex ways that simple surface patterns are not flexible enough to extract them.

The aim of chapter 3 was to find more robust techniques for answer extraction. We compared extraction techniques based on surface patterns with extraction techniques based on dependency patterns. We defined two parallel sets of patterns, one based on surface structures, the other based on dependency relations. These sets of patterns were used to extract and collect facts. The results of the experiments overall showed that the use of dependency patterns indeed has a positive effect, both on the performance of the extraction task as well as on the question answering task. More facts are being extracted with even higher precision. We can conclude that dependency relations eliminate many sources of variation that systems based on surface strings have to deal with.

Still, it is also true that the same semantic relation can sometimes be expressed by several dependency patterns. To account for this syntactic variation we implemented equivalence rules over dependency relations. Using these equivalence rules recall increased a lot. For each relation, the number of extracted facts could have been in-

creased by a similar amount by expanding the number of patterns for that relation. The interesting point is that in this case this was achieved by adding a single, generic component.

Furthermore, we introduced the `d-score`, which computes the extent to which the dependency structure of question and answer match, so as to take into account crucial modifiers that otherwise would be ignored, resulting in more questions being answered by Joost.

The purpose of chapter 4 was to address the low coverage problem of off-line answer extraction by incorporating a state-of-the-art coreference resolution system into an answer extraction system. We investigated the extent to which coreference information could improve the performance of answer extraction, in particular concentrating on the effects of recall. Our aim was to determine whether adding coreference information helped to acquire more facts and to investigate if more questions were answered correctly.

To this end, we built a state-of-the-art coreference resolution system for Dutch, which we evaluated using the MUC score, a clear and intuitive score for evaluating coreference resolution systems. We showed that the results obtained on the KNACK-2002 data were slightly better than the results reported by Hoste (2005), the only other results reported to date on the KNACK-2002 data.

The system has been integrated as an coreference resolution system in the answer extraction module. We evaluated the extended module on the extraction task as well as on the task of question answering (QA). We extracted around 14.5% more facts using coreference based patterns without loss of precision.

Using the extended tables on 233 CLEF questions resulted in an improvement of the performance of a state-of-the-art QA system: 6 more question were answered correctly, which corresponds to an error reduction of 7.3%.

Some questions in the data set seem to be re-formulations of sentences in the newspaper corpus. In a real life application where questions are truly independent of the document collection, adding coreference information would perhaps achieve even better results.

In chapter 5 we automatically learnt patterns to extract facts for off-line question answering by applying a bootstrapping technique. By automatically learning patterns we can more easily adapt the technique of off-line answer extraction to new domains. In addition, we might discover patterns that we ourselves had not thought of. A crucial issue for existing bootstrapping-based algorithms is to ascertain that the patterns and instances found are reliable so as to avoid the extraction of too much noise during

further iterations. It seems natural to focus on precision when working with bootstrapping techniques. Many unreliable patterns will result in noisy sets of instances, which in return will yield wrong patterns, which will extract false instances, etc.

However, storing the facts off-line lets us use frequency counts to filter out incorrect facts afterwards. Then we do not need to focus on precision during the extraction process. It is of greater importance that we extract at least the correct answer.

We presented a bootstrapping algorithm for finding dependency-based patterns and extracting answers. The algorithm takes as input ten seed facts of a particular question category. On the basis of these ten seed pairs patterns are learned. We use these patterns to find new facts and these new facts can again be used to deduce patterns. The evaluation of the newly found patterns is based on the facts they find. Learning patterns for answer extraction based on bootstrapping techniques is therefore only feasible if the answers for a particular set of questions are formulated in a consistent way.

Experiments were performed on three different question categories: capital, football, and minister. For the capital and football categories the results showed indeed that runs with a high recall and low precision achieved the best performance for QA, although for the capital category low precision was more harmful than for the football category. For the minister category one highly precise pattern was on its own responsible for the extraction of most of the facts, so the results for all experiments in this category were the same. We conclude that the results of the experiments suggest that for the benefit of off-line QA we better not focus on high precision and that recall is at least equally important. However, the balance between precision and recall can differ over different categories, depending on the kind of patterns discovered.

Taken together, the results of these experiments suggest that syntactic information can play a crucial role for increasing the recall in off-line answer extraction, not only by defining patterns based on dependency relations, but also in providing useful knowledge to perform coreference resolution for answer extraction. Furthermore, patterns based on dependency relations are easy to learn, since you can simply search for the shortest path between two seed nodes. Surface patterns are typically harder to define in a natural and meaningful way.

In general, off-line answer extraction is a valuable addition to IR-based QA. While IR-based QA is inherently more robust due to the fact that the off-line method is only suitable for questions with clearly defined question terms and answer terms, within the domain of the off-line module scores can be very high for both recall and precision.

6.2 Future work

We conclude this chapter with some ideas for future work.

In this thesis we focused on using syntactic information to increase the recall of off-line answer extraction. Nevertheless, semantic information could also be taken into consideration. We already use lists of terms (e.g., a list of function terms) to make patterns more precise, but similar kinds of information could be used for improving the coverage of patterns. Paşca et al. (2006) show for example how to generalise surface patterns by replacing the terms in the patterns with their corresponding classes of distributional similar words. This technique could also be applied for dependency-based patterns.

With regard to the domain of questions, we only tried to answer factoid CLEF questions in this work. However, by creating off-line complete databases of facts other kinds of questions could also be considered, the most obvious being list questions. For instance, a query like "Name five presidents of the United States." could easily be looked up in the table of function facts. Since 2006 the question sets provided by CLEF include list questions.

It would also be interesting to see what kind of questions users would ask given an online question answering system based on newspaper text and if they are suitable for off-line answer extraction. If we could daily create a database with tables of facts for a newspaper published in the morning, the advantage would be that users can ask questions about very recent events (e.g. scores for sport matches of the previous day, new books being published, people who died) or even about future events (e.g. questions about the weather). Moreover, this technique of off-line answer extraction typically makes the time a user has to wait for an answer much shorter than for IR-based question answering.

Using an online setting and real user questions would also give better insights into the potential of coreference resolution for question answering. The CLEF questions we used gave us the impression that they were not created completely independent of the corpus.

In this thesis, we only used the coreference resolution system for off-line question answering. Of course it can also be deployed in other parts of the question answering system, for example in the answer extraction phase after passage retrieval. Starting from 2007 topic questions were introduced in the question set. Topic questions are clusters of questions which are related to the same topic and possibly contain anaphoric references between one question and the other questions. A reference resolution system

is needed in these cases.

We describe another possibility to use coreference knowledge for question answering in Tiedemann and Mur (to appear). We performed experiments were we segmented CLEF documents into passages based on coreference chains. Compared to the baseline where documents were segmented into paragraphs, defined by the paragraph markup of the CLEF corpus itself, we improved the scores for QA. However, from the same experiments it turned out that paragraphs of fixed sizes were to be preferred, while the paragraphs based on coreference chains varied a lot in length. It remains to be seen what can be achieved by using coreference chains resulting in paragraphs of more or less the same size.